PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-155025
(43)Date of publication of application : 08.06.2001

(51)Int.Cl. G06F 17/30

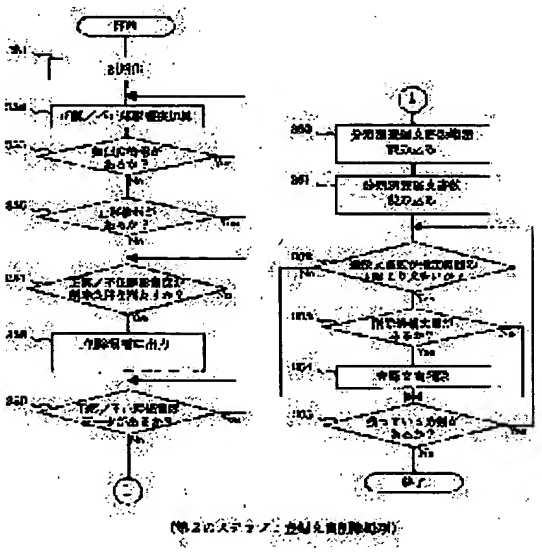
(21)Application number : 11-335442
(22)Date of filing : 26.11.1999

(71)Applicant : TOSHIBA CORP
TOSHIBA COMPUT ENG CORP

(72)Inventor : NAKAZATO SHIGEMI
SAITO HIROMI
KOBAYASHI TSUTOMU
MATSUKUMA TAKESHI
NAKAMOTO YUKIO
NISHINA TAKUYA
YAMAZAKI HIROSHI

(54) DOCUMENT SORTING DEVICE AND METHOD, AND DATABASE UPDATING METHOD

(57)Abstract:
PROBLEM TO BE SOLVED: To solve a problem that a conventional document sorting device does not consider efficient updating of a database to be used for sorting document for preventing lowering in the retrieval accuracy of a document.
SOLUTION: In the database to be used for the document sorting device, similarity between document data and registered document data in the database is calculated and when sorting of the registered document data is matched with the sorting of the document data, this calculated similarity is stored as a right answer influence degree but when sorting of the said registered document data is not matched with the sorting of the document data, the similarity is stored as a wrong answer influence degree. Since the database is updated so as to delete the registered document data from the database by sorting corresponding to the right answer influence degree and wrong answer influence degree, even when a word to be used for the same sorting in the database is changed with the passage of time, appropriate sorting can be applied without increasing the number of registered documents needlessly.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号
特開2001－155025
（P2001－155025A）
(43)公開日　平成13年６月８日(2001.6.8)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード(参考)
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/40	３７０Ａ　５Ｂ０７５
			15/401	３１０Ｄ

審査請求　未請求　請求項の数12　ＯＬ　（全 11 頁）

(21)出願番号　特願平11－335442

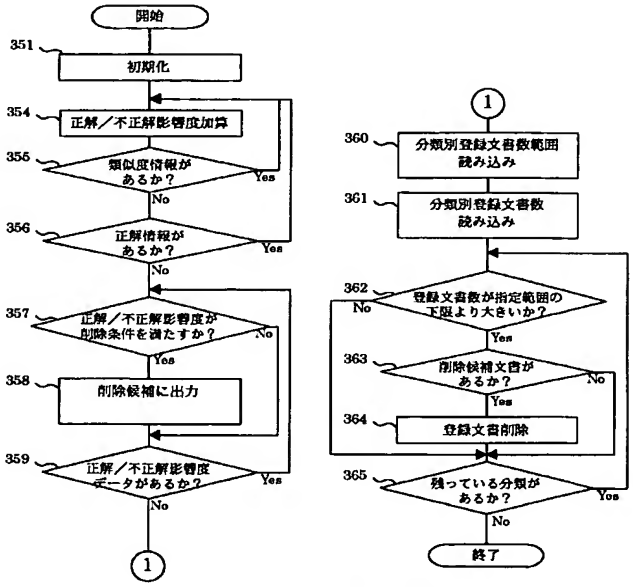
(22)出願日　平成11年11月26日(1999.11.26)

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(71)出願人　000221052
東芝コンピュータエンジニアリング株式会社
東京都青梅市新町３丁目３番地の１
(72)発明者　中里　茂美
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内
(74)代理人　100083161
弁理士　外川　英明

最終頁に続く

(54)【発明の名称】　文書分類装置、文書分類方法、及びデータベース更新方法

(57)【要約】
【課題】　従来の文書分類装置においては、文書分類のために用いるデータベースを文書の検索精度を低下させないよう効率的に更新することについて考慮されていなかった。
【解決手段】　文書分類装置に用いるデータベースに関し、文書データとデータベース内の登録文書データとの類似度を算出し、この算出された類似度を、上記登録文書データの分類と上記文書データの分類とが一致している場合正解影響度とし、上記登録文書データの分類と上記文書データの分類とが不一致の場合不正解影響度として記憶し、上記正解影響度及び不正解影響度に応じて登録文書データの分類別に上記データベースから削除するようにデータベースの更新を行なうようにしたため、登録された文書の数を不必要に増加させずにデータベース中の同一分類の中で使用される単語に時間の経過に伴う変化が生じる場合でも適切な分類付与を行なうことができるようになる。



（第２のステップ：登録文書削除処理）

1

【特許請求の範囲】

【請求項１】予め分類が付与された複数の登録文書データを格納するデータベースと、このデータベースに含まれない文書データを入力する第１の入力手段と、この入力された文書データと上記データベース内の登録文書データとの類似度を算出する算出手段と、この算出手段にて算出された類似度を出力する第１の出力手段と、上記算出手段で求められた類似度を、上記文書データに付与されるべき正解分類に一致すれば正解影響度とし、上記正解分類に不一致であれば不正解影響度として記憶する記憶手段と、この記憶手段に記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべきものを出力する第２の出力手段と、を具備することを特徴とする文書分類装置。

【請求項２】予め分類が付与された複数の登録文書データを格納するデータベースと、このデータベースに含まれない文書データを入力する第１の入力手段と、この入力された文書データと上記データベース内の登録文書データとの類似度を算出する算出手段と、この算出手段で求められた類似度に基づき上記文書データに付与されるべき分類を正解分類として入力する第２の入力手段と、この第２の入力手段により入力された上記正解分類と同じ分類が付与された登録文書データの類似度は正解影響度とし、上記正解分類と異なる分類が付与された登録文書データの類似度は不正解影響度として記憶する記憶手段と、この記憶手段に記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべきものを出力する出力手段と、を具備することを特徴とする文書分類装置。

【請求項３】予め分類が付与された複数の登録文書データを格納するデータベースと、このデータベースに含まれない文書データを入力する第１の入力手段と、この入力された文書データと上記データベース内の登録文書データとの類似度を算出する算出手段と、この算出手段で求められた類似度に基づき上記文書データに付与されるべき分類を正解分類として入力する第２の入力手段と、この第２の入力手段により入力された上記正解分類と同じ分類が付与された登録文書データの類似度は正解影響度とし、上記正解分類と異なる分類が付与された登録文書データの類似度は不正解影響度として記憶する記憶手段と、この記憶手段に記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべき登録文書データを選択する選択手段と、この選択手段にて選択された登録文書データを上記データベースから削除する削除手段と、を具備することを特徴とする文書分類装置。

【請求項４】上記選択手段は、上記正解影響度が第１のしきい値より小さく、且つ上記不正解影響度が第２のしきい値より大きい場合、削除すべきものとして選択することを特徴とする請求項３記載の文書分類装置。

2

【請求項５】上記選択手段は、上記正解影響度と上記不正解影響度の相対的な関係から削除すべきものとして選択することを特徴とする請求項３記載の文書分類装置。

【請求項６】上記選択手段は、上記正解影響度及び上記不正解影響度、上記正解影響度と上記不正解影響度とのそれぞれの回数、及び登録文書の上記データベースへの登録からの経過時間に基づいて削除すべきものとして選択することを特徴とする請求項３記載の文書分類装置。

【請求項７】予め分類が付与された複数の登録文書データを格納するデータベースと、このデータベースに含まれない文書データを入力する第１の入力手段と、この入力された文書データと上記データベース内の登録文書データとの類似度を算出する算出手段と、この算出手段で求められた類似度に基づき上記文書データに付与されるべき分類を正解分類として入力する第２の入力手段と、この第２の入力手段により入力された上記正解分類と同じ分類が付与された登録文書データの類似度は正解影響度とし、上記正解分類と異なる分類が付与された登録文書データの類似度は不正解影響度として記憶する記憶手段と、上記データベースに保持される分類別の登録文書数の範囲を入力する第３の入力手段と、この第３の入力手段にて入力された分類別の登録文書数の範囲を超える場合、この登録文書数の範囲に収まるように上記記憶手段に記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべき登録文書データを選択する選択手段と、この選択手段にて選択された登録文書データを上記データベースから削除する削除手段と、を具備することを特徴とする文書分類装置。

【請求項８】予め分類が付与された複数の登録文書データを格納するデータベースを有し、このデータベースに格納された登録文書データとの類似度に基き所定の文書データに分類を付与する文書分類方法において、文書データと上記データベース内の登録文書データとの類似度を算出し、この算出された類似度に基き文書データを分類し、上記算出された類似度を、上記文書データに付与されるべき正解分類に一致すれば正解影響度とし、上記正解分類に不一致であれば不正解影響度として記憶し、この記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべきものを出力することを特徴とする文書分類方法。

【請求項９】予め分類が付与された複数の登録文書データを格納するデータベースを有し、このデータベースに格納された登録文書データとの類似度に基き所定の文書データに分類を付与する文書分類方法において、文書データと上記データベース内の登録文書データとの類似度を算出し、この算出された類似度に基き文書データを分類し、上記算出された類似度を、上記文書データに付与されるべき正解分類に一致すれば正解影響度とし、上記正解分類に不一致であれば不正解影響度として記憶し、

この記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべきものを出力し、削除すべきものとして出力された登録文書データを削除して、上記データベースを更新することを特徴とする文書分類方法。

【請求項１０】予め分類が付与された複数の登録文書データを格納するデータベースを有し、このデータベースに格納された登録文書データとの類似度に基き所定の文書データの分類を付与する文書分類方法において、文書データと上記データベース内の登録文書データとの類似度を算出し、この算出された類似度に基き文書データを分類し、上記データベースにおける登録文書データ数の範囲を分類毎に設定し、上記算出された類似度を、上記文書データに付与されるべき正解分類に一致すれば正解影響度とし、上記正解分類に不一致であれば不正解影響度として記憶し、上記登録文書データ数の範囲を超える場合、この範囲に収まるように上記正解影響度及び不正解影響度に応じて上記データベースから削除すべきものを出力し、削除すべきものとして出力された登録文書データを削除して、上記データベースを更新することを特徴とする文書分類方法。

【請求項１１】所定の文書データに類似する登録文書データを抽出するために用いられ、予め分類が付与された複数の登録文書データを格納するデータベースを更新するデータベース更新方法において、文書データと上記データベース内の登録文書データとの類似度を算出し、この算出された類似度を、上記登録文書データの分類と上記文書データの分類とが一致している場合正解影響度とし、上記登録文書データの分類と上記文書データの分類とが不一致の場合不正解影響度として記憶し、上記正解影響度及び不正解影響度に応じて登録文書データの分類別に上記データベースから削除することを特徴とするデータベース更新方法。

【請求項１２】所定の文書データに類似する登録文書データを抽出するために用いられ、予め分類が付与された複数の登録文書データを格納するデータベースを更新するデータベース更新方法において、上記データベースにおける登録文書データ数の範囲を分類毎に設定し、文書データと上記データベース内の登録文書データとの類似度を算出し、この算出された類似度を、上記登録文書データの分類と上記文書データの分類とが一致している場合正解影響度とし、上記登録文書データの分類と上記文書データの分類とが不一致の場合不正解影響度として記憶し、上記登録文書データ数の範囲を超える場合、この範囲に収まるように上記正解影響度及び不正解影響度に応じて上記データベースから削除することを特徴とするデータベース更新方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】　本発明は、電子化され分類

登録された複数文書のデータベースを用いて、新規の文書データがどの分類に対応するかを求めるための文書分類装置及び文書分類方法、及び上記データベースを更新する方法に関する。

【０００２】

【従来の技術】　近年大量の電子化された文書データが流通するようになり、各文書データがどのような分類に属するかを自動的に判別する技術が実用化されている。一般的な技術としては、入力された文書と登録されている文書間の類似度を共通単語数やベクトル空間法等を用いて求め、類似度の高い登録文書が属する分類に基づいて、入力文書の分類を特定するという方法が知られている。例えば、本願発明者らが先に特許出願した特願平１１－１２００２３号の明細書及び図面に記載されているように、登録文書が属する分類と類似度とを利用して分類間における類似度の正解率の違いを反映して類似文書を検索するものがあり、登録文書毎に分類が特定されているので、検索により抽出された文書から入力文書の分類を特定することができる。

【０００３】

【発明が解決しようとする課題】　従来の装置においては、類似文書の検索結果により入力文書の分類を特定する手法を用いているが、時間経過につれて技術が進歩し使用される単語が変化していくため、文書の検索精度を低下させないよう効率的にデータベースを更新することについて考慮されていなかった。

【０００４】例えば、パーソナルコンピュータのような技術進歩の激しい分類において、ＣＰＵ及びＯＳの名称や、音声及び画像のフォーマット等新しい単語が次々と現れて来るため、これらの単語を含む新しい文書を登録して行かなければ、対象文書に対応する文書を検索することができなくなり、対象文書を検索された文書と同じ分類を付与すると適当でない場合が増加する。

【０００５】このような事態を回避すべく新文書を次々と登録して単純に各分類の代表となる文書の登録数を増やすことが考えられるが、これはデータベースが肥大化し、分類速度の低下及びメモリや外部記憶装置等のリソース不足等の別な問題を招くことになるため好ましくない。よってデータベースの更新の際には文書の登録だけでなく不要な文書の削除を行なう必要が有る。

【０００６】しかしながらデータベースに既に登録された文書を例えば単に古いものから削除すると、その分類の代表的なものを削ってしまう場合もあり好ましくない。別な手法として人手により内容確認しつつ削除することも考えられるが、これでは手間が掛かるばかりか客観的に削除すべきものか否かの判断ができないという問題がある。

【０００７】本発明は、このような課題を解決するためのもので、同一分類の中で使用される単語に時間の経過に伴う変化が生じる場合でも適切な分類付与を行なうこ

5

とのできる文書分類装置及び文書分類方法を提供することを目的とする。

【０００８】また本発明は、文書分類装置に使用されるデータベースについて、登録された文書の数を不必要に増加させずにデータベースの更新を行なうデータベース更新方法を提供することを目的とする。

【０００９】

【課題を解決するための手段】　上記目的を達成するために、請求項１に係る発明では、予め分類が付与された複数の登録文書データを格納するデータベースと、このデータベースに含まれない文書データを入力する第１の入力手段と、この入力された文書データと上記データベース内の登録文書データとの類似度を算出する算出手段と、この算出手段にて算出された類似度を出力する第１の出力手段と、上記算出手段で求められた類似度を、上記文書データに付与されるべき正解分類に一致すれば正解影響度とし、上記正解分類に不一致であれば不正解影響度として記憶する記憶手段と、この記憶手段に記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべきものを出力する第２の出力手段と、を具備することを特徴とする。このような構成により、同一分類の中で使用される単語に時間の経過に伴う変化が生じる場合でも、データベースから不要な登録文書データを削除し適当な登録文書データを残すことができるので、適切な分類付与を行なうことができる。

【００１０】また、本発明の文書分類装置は請求項２に記載されるように、予め分類が付与された複数の登録文書データを格納するデータベースと、このデータベースに含まれない文書データを入力する第１の入力手段と、この入力された文書データと上記データベース内の登録文書データとの類似度を算出する算出手段と、この算出手段で求められた類似度に基づき上記文書データに付与されるべき分類を正解分類として入力する第２の入力手段と、この第２の入力手段により入力された上記正解分類と同じ分類が付与された登録文書データの類似度は正解影響度とし、上記正解分類と異なる分類が付与された登録文書データの類似度は不正解影響度として記憶する記憶手段と、この記憶手段に記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべき登録文書データを選択する選択手段と、この選択手段にて選択された登録文書データを上記データベースから削除する削除手段と、を具備することを特徴とする。このような構成とすることにより、同一分類の中で使用される単語に時間の経過に伴う変化が生じる場合でも、データベースから不要な登録文書データを削除し適当な登録文書データを残すことができるので、適切な分類付与を行なうことができる。

【００１１】また、本発明の文書分類装置は請求項７に記載されるように、予め分類が付与された複数の登録文書データを格納するデータベースと、このデータベースに含まれない文書データを入力する第１の入力手段と、この入力された文書データと上記データベース内の登録文書データとの類似度を算出する算出手段と、この算出手段で求められた類似度に基づき上記文書データに付与されるべき分類を正解分類として入力する第２の入力手段と、この第２の入力手段により入力された上記正解分類と同じ分類が付与された登録文書データの類似度は正解影響度とし、上記正解分類と異なる分類が付与された登録文書データの類似度は不正解影響度として記憶する記憶手段と、上記データベースに保持される分類別の登録文書数の範囲を入力する第３の入力手段と、この第３の入力手段にて入力された分類別の登録文書数の範囲を超える場合、この登録文書数の範囲に収まるように上記記憶手段に記憶された各登録文書データの上記正解影響度及び不正解影響度に応じて上記データベースから削除すべき登録文書データを選択する選択手段と、この選択手段にて選択された登録文書データを上記データベースから削除する削除手段と、を具備することを特徴とする。このような構成とすることにより、同一分類の中で使用される単語に時間の経過に伴う変化が生じる場合でも、適当な数の登録文書データを残すことができ、適切な分類付与を行なうことができる。

【００１２】また、本発明のデータベース更新方法は請求項１１に記載されるように、所定の文書データに類似する登録文書データを抽出するために用いられ、予め分類が付与された複数の登録文書データを格納するデータベースを更新するデータベース更新方法において、文書データと上記データベース内の登録文書データとの類似度を算出し、この算出された類似度を、上記登録文書データの分類と上記文書データの分類とが一致している場合正解影響度とし、上記登録文書データの分類と上記文書データの分類とが不一致の場合不正解影響度として記憶し、上記正解影響度及び不正解影響度に応じて登録文書データの分類別に上記データベースから削除することを特徴とする。このような構成とすることにより、登録された文書の数を不必要に増加させずにデータベースの更新を行なうことができる。

【００１３】

【発明の実施の形態】以下、図面を参照して本発明の一実施形態を説明する。

【００１４】図１は本発明の一実施形態に係る文書分類装置のハードウェア構成を示す図である。なお、本装置は一般的なアーキテクチャを持つコンピュータ上の一機能として構築されるものである。

【００１５】図１に示すように、本装置は、制御装置１、キーボード、ポインティングデバイス、スキャナを有する入力装置２、類似文書の検索結果などを表示する表示装置３、および外部記憶装置４から構成される。この外部記憶装置４は、例えばハードディスク装置、また

はＤＶＤ（Ｄｉｇｉｔａｌ　Ｖｉｄｅｏ　Ｄｉｓｃ）装置などからなる。

【００１６】図２に本装置における制御装置１の構成を示す。制御装置１はＣＰＵ、ＲＯＭ、及びＲＡＭを有しており、図２の中ではＣＰＵとＲＯＭにより為される部分を機能的にプログラム部２００とし、ＲＡＭを機能的にバッファ部２５０として表わしている。

【００１７】プログラム部２００は、初期化部２０１、分類文書入力部２０２、登録文書読み込み部２０３、類似度算出部２０４、分類特定部２０５、正解分類情報入力部２０６、正解／不正解影響度算出部２０７、削除候補判断条件設定部２０８、影響度判定・削除候補出力部２０９、削除文書選択部２１０、登録文書削除部２１１、分野別登録文書数範囲設定部２１２、及び分野別登録文書数読み込み部２１３の１３の機能を有している。

【００１８】バッファ部２５０は、分類文書格納バッファ部２５１、登録文書格納バッファ部２５２、類似度算出結果格納バッファ部２５３、正解分類情報格納バッファ部２５４、正解／不正解影響度格納バッファ部２５５、削除候補判断条件格納バッファ部２５６、削除候補格納バッファ部２５７、分類別登録文書数範囲格納バッファ部２５８、及び分類別登録文書数格納バッファ部２５９の９の領域を有している。

【００１９】初期化部２０１は、バッファ部２５０内の各バッファ部のデータのクリアを行う。

【００２０】分類文書入力部２０２は、ユーザが入力装置２を用いて入力する分類文書データを、分類文書格納バッファ部２５１へ格納する。この時、分類文書入力部２０２は分類文書ＩＤを発行し、この分類文書ＩＤも分類文書データと共に分類文書格納バッファ部２５１へ格納する。

【００２１】登録文書読み込み部２０３は、外部記憶装置4に格納された登録文書を読み出し、登録文書格納バッファ部２５２へ格納する。

【００２２】類似度算出部２０４は、分類文書格納バッファ部２５１に格納されている分類文書と、登録文書格納バッファ部２５２に格納されている登録文書とを単語に分割し、各単語の出現回数をベクトルの成分とするベクトル空間法を用いて両文書の類似度を算出する。さらに分類文書ＩＤ、登録文書ＩＤ、類似度、及び登録文書が属する分野情報を組にして、類似度算出結果格納バッファ部２５３に格納する。尚、ここで類似度はベクトル空間法を用いる代わりに共通単語数を用いて算出しても構わない。

【００２３】分類特定部２０５は、類似度算出結果格納バッファ部２５３に格納されている登録文書との類似度情報から、各分類別に類似度を加算した分類ー類似度一覧表を作成し、類似度の和の大きい物から順に分類特定結果として出力する。尚、上述のように類似度加算の一覧表により分類を特定する方法でなく、類似度の高い文書の属する分類をそのまま利用する方法でも構わない。

【００２４】正解分類情報入力部２０６は、入力文書の分類後ユーザの入力装置２を介して入力される当該入力文書に対して正解となる分類の情報と分類文書ＩＤとを受け付けて、正解分類情報格納バッファ部２５４へ格納する。

【００２５】正解／不正解影響度算出部２０７は、類似度算出結果格納バッファ２５３に格納されている類似度算出結果情報と、正解分類情報格納バッファ部２５４に格納された分類文書の正解分類情報から正解／不正解影響度を算出する。その算出方法は、登録文書別に、分類文書の正解分類と同じ分類に属する場合はその登録文書の類似度を正解影響度に、一方正解分類と異なる分類に属する場合はその登録文書の類似度を不正解影響度に、それぞれ加算した後正解／不正解影響度格納バッファ部２５６に格納する。尚、登録文書が複数の分類に属する場合や、正解分類が複数ある場合は、そのすべてについて正解／不正解影響度を算出しても構わないし、その中の代表のみについて正解／不正解影響度を算出しても構わない。

【００２６】削除候補判断条件設定部２０８は、ユーザが入力装置２より入力した削除候補判断条件を削除候補判断条件格納バッファ部２５６に格納する。

【００２７】影響度判定・削除候補出力部２０９は、正解／不正解影響度格納バッファ部２５５に格納されている登録文書ごとの正解／不正解影響度と、削除候補判断条件格納バッファ部２５６に格納されている削除候補判断条件から、正解／不正解影響度の値が条件を満たす文書を選択し、削除候補格納バッファ部２５７に格納する。

【００２８】削除文書選択部２１０は、削除候補格納バッファ部２５７に格納されている削除候補文書情報と、分類別登録文書数範囲格納バッファ部２５８に格納されている分類別登録文書数範囲情報と、分類別登録文書数格納バッファ部２５９に格納されているデータベースに格納されている分類別の登録文書数の情報から、分類ごとに指定の文書数範囲に収まるように、不正解影響度から正解影響度を引いた値の大きい順に選択し、削除候補格納バッファ部２５７に格納する。尚、選択方法としては、不正解影響度の大きいものから順に選択する方法でも構わないし、登録された日付の古い文書から順に選択しても構わない。

【００２９】登録文書削除部２１１は、削除候補格納バッファ部２５７に格納されている削除候補文書情報を元に、データベースから登録文書を削除する。

【００３０】分類別文書数範囲設定部２１２は、入力装置２によりユーザが入力した登録文書数の上下限の値を、分類別文書数範囲格納バッファ部２５８に設定する。

【００３１】分類別文書数読み込み部２１３は、外部記

憶装置4に格納されているデータベースの分類別登録文書数情報を分類別文書数格納バッファ部２５９に読み込む。

【００３２】次に、本実施形態の文書分類装置の動作を説明する。ここで説明する動作は制御装置１のＣＰＵが、ＲＯＭ内のプログラム、及びＲＡＭ内の記憶領域を用いて実行するものである。

【００３３】本実施形態は、大きく第１のステップと第２のステップとからなる。第１のステップは、文書分類装置に登録された文書から、削除すべき文書を選択するために、文書分類装置に登録されていない文書の分類処理を行い、その処理結果を蓄積するステップである。第２のステップは、この作成された処理結果と、実際にその文書が属する分類（正解分類）の情報をもとに、削除すべき文書を選択し、削除を行うステップである。

【００３４】まず、分類処理結果を蓄積する第１のステップについて説明する。

【００３５】はじめにユーザは、入力装置２を用いて、外部記憶装置４に文書の分類時に参照する文書データを格納する（ステップ３０１）。続いて初期化部２０１により全バッファをクリアする（ステップ３０２）。

【００３６】次に分類文書入力部２０２が、入力装置２を通じてユーザより入力される分類文書を受け付けて、分類文書格納バッファ部２５１にこの文書を検索キー文書として分類文書ＩＤと共に格納する（ステップ３０３）。具体例として、図４に示すような内容の分類文書（本文）を検索キー文書の一つとして分類文書ＩＤ「１」と共に格納したものとする。

【００３７】登録文書読み出し部２０３は、外部記憶装置４に格納された複数の文書を読み出し、登録文書格納バッファ部２５２にこれらの文書を登録文書として格納する（ステップ３０４）。登録文書には、文書を識別するための登録文書ＩＤと、その文書の分類を表す分類情報が付与されている。具体例として、図５に示すような、登録文書ＩＤ、分類、及び本文からなるデータを複数格納したとする。

【００３８】類似度算出部２０４は、分類文書格納バッファ部２５１に格納された分類文書の本文と、登録文書格納バッファ部２５２に格納された登録文書の本文とを比較し両文書の類似度を算出する。算出された類似度は登録文書ＩＤ及びその登録文書の分類を表す分類情報と共に、類似度算出結果格納バッファ部２５３に格納する（ステップ３０５）。この時、類似度が大きいものから一定の件数だけ格納したり、一定の類似度以上のものだけを格納しても構わない。図６に示す類似度算出結果格納例では、最上段に文書ＩＤ＝１０２３、分類=テレビ、類似度=０．３７８という内容が格納されており、このような情報が分類文書ＩＤ＝１に対する分類毎の類似度の大きさとして複数あることを示している。

【００３９】類似度の格納が済むと、類似度を算出していない登録文書が残っているかを判断し（ステップ３０６）、残っている場合は、ステップ３０４に戻って残りの登録文書に対してステップ３０４、及び３０５の動作を行う。一方、他に登録文書が無いと判断した場合は、ステップ３０７に進む。

【００４０】ステップ３０７では、ステップ３０５で類似度算出結果格納バッファ部２５３に格納した類似度算出結果の登録文書ごとの類似度を、その登録文書の属する分類別に集計し、分類－類似度一覧表を作成する（ステップ３０７）。図７の一覧表の例では、パソコン分野に属する文書の類似度の和が１．７８２であり、ビデオ分野に属する文書の類似度の和が１．０２３であることを示している。

【００４１】ステップ３０８では、作成した分類－類似度一覧表を分類毎の類似度の和でソートし、値の大きいものから順に分類結果として出力する。分類結果が出力されると、ユーザは分類文書に対して高い類似度の分類を付与するべく、入力装置２から分類を入力する（ステップ３０９）。ここで入力された分類は分類文書に対する正解分類として、正解分類情報入力部２０６を介して、分類文書のＩＤと共に正解分類情報格納バッファ部２５４に格納される（ステップ３５２）。ここで正解分類とは、ユーザが図３のステップ３０９の後に付与した分類文書毎の最も高い類似度の分類である。図８は、分類文書ＩＤ＝１の文書の正解分類「パソコン」が入力された場合の格納例である。尚、ユーザによる分類付与に際しては、最も高い類似度のみでなく、上から２位まで又は３位までを付与するようにしても構わない。

【００４２】分類と共に正解分類情報格納バッファ２５４に格納された分類文書は、自動発番される登録文書ＩＤと組にしたデータとして外部記憶装置４へ格納され、別途登録文書として使用される。

【００４３】分類付与が済むと、分類文書格納バッファ２５１に未分類の分類文書が残っているかを判断し、残っていればステップ３０３に戻り、ステップ３０３から３０８までを繰り返す。一方、分類文書が残っていなければ分類処理を終了する（ステップ３１０）。

【００４４】次に、第１のステップで格納した類似度算出結果及び正解分類を用いて、登録文書の一部を削除する第２のステップについて説明する。図８は、その手順を示すフローチャートである。

【００４５】はじめに初期化部２０１により類似度度算出結果格納バッファ部２５３以外のバッファをクリアし、削除候補判断条件設定部２０８により削除候補判断条件格納バッファ部２５６に格納する（ステップ３５１）。

【００４６】次に、ステップ３０５で類似度算出結果格納バッファ部２５３に格納した類似度算出結果と、正解分類情報格納バッファ部２５４に格納した分類文書ＩＤとその正解分類情報とに基づき、正解の分類と不正解の

分類とに応じて類似度を正解／不正解影響度格納バッファ部２５５の所定の領域に加算する（ステップ３５４）。分類文書ＩＤが１で、正解分類がパソコンであった場合、類似度算出結果が図６の状態であるとすると、登録文書ＩＤ=１０２３の文書は、その分類が正解分類と異なるので、その類似度０．３７８を不正解影響度の領域に加算する。登録文書ＩＤ＝９９２４の文書は、その分類が正解分類と一致するので、その類似度０．２２６を正解影響度の領域に加算する。以下の類似度算出結果についても同様に処理する。正解／不正解影響度格納バッファ部２５５の例を図１０に示す。登録文書ＩＤ＝１の文書は、分類が「家具」で、家具が正解の分類文書との類似度の和が０．５６２８５であり、「家具」以外の分類の分類文書との類似度の和が０．００８４５であることを表している。尚、正解、及び不正解の各影響度は文書間の類似度の和をとるため、１以上の値をとり得る。

【００４７】この後、処理中の分類文書の類似度算出結果が残っているか判断し（ステップ３５５）、残っている場合はステップ３５４に戻り、ステップ３５４の処理を繰り返す。この処理の対象となるのはすべての類似度算出結果でも構わないし、類似度の高いものから所定件数分、または類似度が一定の値以上のものでも構わない。一方、処理する類似度算出結果が残っていない場合はステップ３５６に進む。

【００４８】ステップ３５６では、正解情報が残っているかを判断し、残っている場合はステップ３５２に戻り、上述したステップ３５２から３５５までの処理を繰り返し、残っていなければ、ステップ３５７に進む。

【００４９】ステップ３５７では、影響度判断・削除候補出力部２０９が、ステップ３５１で削除候補判断条件格納バッファ部２５６に格納した削除候補判断条件をもとに、正解／不正解影響度格納バッファ部２５５に格納されている正解／不正解影響度をチェックし、条件を満たしていればステップ３５８に進み、満たしていなければステップ３５９に進む。削除候補判断条件格納バッファ部２５６が図１１の状態で、正解／不正解影響度格納バッファ部２５５が図１０の状態である場合、条件を満たすのは、登録文書ＩＤが２の文書と同じく９９２４の文書とである。

【００５０】ステップ３５８では、影響度判定・削除候補出力部２０９が、ステップ３５７で条件を満たした文書の、登録文書ＩＤ、分類情報、正解影響度、及び不正解影響度の各データを削除候補格納バッファ部２５７に格納する。図１２は、図１０の正解不正解影響度情報から図１１の正解／不正解影響度条件、すなわち正解影響度が０．０１未満、且つ不正解影響度が０．２０以上を満たす文書情報を格納した状態である。

【００５１】続いて正解／不正解影響度データが残っているかを判断し（ステップ３５９）、残っていればステップ３５７に戻り、ステップ３５７から３５８を繰り返す。正解／不正解影響度データが残っていない場合はステップ３６０に進む。

【００５２】ステップ３６０では、分類別登録文書数範囲設定部２１１が、入力装置２よりユーザが入力した分類別登録文書数の上下限値を、分類別登録文書数範囲格納バッファ部２５８に格納する。図１３は分類別登録文書数範囲の格納例である。この例では、パソコン分類の登録文書は、全体文書数の３±０．５％が指定されていることを示す。このような比率の範囲を指定する以外に、分野別登録文書の件数の範囲を指定しても構わない。

【００５３】ステップ３６１では、分類別登録文書数読み込み部２１２が、分類別登録文書数格納バッファ部２５９に全登録文書数と分類別の登録文書数を読み込む。図１４は分類別登録文書数の格納例である。この例では、全文書数が２８０６件、パソコン分類の登録文書数が９０件で全文書数に対する比率が３．２％、インターネット分類の登録文書数が１１９件で全文書数に対する比率が４．２％であることを示している。

【００５４】ステップ３６２では、分類別登録文書数範囲格納バッファ部２５８に格納されている分類別登録文書数範囲情報と分野別登録文書数格納バッファ部２５９に格納されている登録文書数情報を比較する。各分野毎の登録文書数が指定の範囲の下限より大きい場合は、ステップ３６３に進み、小さい場合はステップ３６５に進む。分類別登録文書数範囲格納バッファが図１３の状態にあり、分類別登録文書数格納バッファが図１４の状態にある場合、パソコン分類の文書数の比率３．２％は分布範囲の下限である２．５％よりも大きいので条件を満たしている。一方インターネット分類の文書数の比率４．２％は、分布範囲の下限である４．５％よりも小さいので条件を満たしていないことになる。

【００５５】ステップ３６３では、その分類の削除文書候補があるかを判断し、ある場合はステップ３６４に進み、無い場合はステップ３６５へ進む。ステップ３６４では、その分野の登録文書数が、指定範囲の下限を下回らない範囲で削除する。従ってパソコン分類は分布範囲内であり、且つ図１２に示す通り削除候補格納バッファ部２５７にパソコン分類の候補として登録ＩＤが９９２４の登録文書が登録されているので、この登録ＩＤが９９２４の登録文書が削除されることになる。

【００５６】ステップ３６５では、残っている分類があるかを判断し、あればステップ３６２に戻りステップ３６２から３４６までを繰り返し、無ければ処理を終了する。

【００５７】上述の通り、分類文書との類似度を求める度に、各登録文書毎に正解影響度及び不正解影響度のデータを蓄積していくので、外部記憶装置４内に構築されたデータベースから登録文書の削除を行なう際に削除す

べき登録文書、残すべき登録文書とを客観的に把握することができる。よって、データベース内の適正件数を保持しつつそれを超える場合に文書分類に悪影響を及ぼす登録文書から削除することができ、結果として文書分類の際に類似度の高い登録文書の分類が正しい分類となるように分類付与することができるようになる。

【００５８】尚、上記実施形態では正解分類の情報をユーザが入力するように構成したが、分類文書と共にデータとして予め格納しておいても構わない。この場合、主たる目的はデータベース更新に使用することとする。登録文書の分類及び類似度と上記正解分類とから削除候補抽出を行なう点については、上記実施形態と同様である。

【００５９】また、上記実施形態では、正解影響度条件及び不正解影響度条件を単純なしきい値として設定したが、正解影響度と不正解影響度の相対的な関係、それぞれの回数、登録または文書作成からの経過時間を加味し、これらの要因を踏まえ、以下の計算式を用いてより適切な削除候補を求めることもできる。

【００６０】ｉ）不正解影響度が０でない場合、

Ｃ＝（Ｂ×Ｂ／（Ｇ＋ｘ））／（ＢＫ＋ｙ）＋Ｄ×ｚ

ｉｉ）不正解影響度が０の場合、

Ｃ＝－（Ｇ×Ｇ／（Ｂ＋ｘ））／（ＧＫ＋ｙ）＋Ｄ×ｚ

但し、Ｃ　：削除対象度合い（値が大きい方がより削除すべき）、

Ｂ　：不正解影響度、

Ｇ　：正解影響度、

ＢＫ：不正解影響回数、

ＧＫ：正解影響回数、

Ｄ　：登録からの経過時間（日数、又は月数等）

ｘ　：定数１

ｙ　：定数２

ｚ　：定数３

「ｉｉ）不正解影響度が０の場合」は、通常であれば削除対象とすべき登録文書ではないのであるが、１件以上削除しなければならない状況で全件「ｉｉ）不正解影響度が０」であると順位が付けられなくなるので、この点を考慮して算出するものである。

【００６１】これらの式を用いて削除対象度合いを算出した例を図１５に示す。ここでは各定数をｘ＝０．１，ｙ＝３，ｚ＝０．０１として計算した。この例では時間の経過による影響が比較的小さくなるように定数を設定した。その結果、削除対象度合いの大きい方から登録文書ＩＤは２０５６，６７７２，９９２４という順になる。分布範囲との関係から例えば１件削除すべきという場合には、その最上位の登録文書ＩＤが２０５６のデータを削除することになる。

【００６２】本発明はその主旨を逸脱しない範囲であれば、上記の実施例に限定されるものではない。そして、データベース検索装置、及び文書検索装置等に広く適用できるものである。

【００６３】

【発明の効果】以上詳述したように本発明によれば、各登録文書毎に正解影響度及び不正解影響度を把握し、登録文書を増やすだけでなく不正解影響度の高い登録文書を削除するようにするため、同一分類の中で使用される単語に時間の経過に伴う変化が生じるような場合でも適切な分類付与を行なえる。

【００６４】またデータベースから登録文書の削除を行なう際に削除すべき登録文書、残すべき登録文書とを客観的に把握することができ、データベースの更新を適切に行なうことができる。

【図面の簡単な説明】

【図１】本発明に係る一実施形態の文書分類装置のハードウェア構成を示す図

【図２】図１の文書分類装置における制御装置の機能ブロック図

【図３】文書データと登録文書データの類似度を算出し分類を付与する手順を示す図

【図４】分類文書格納例を示す図

【図５】登録文書格納例を示す図

【図６】類似度算出結果バッファ部格納例を示す図

【図７】分類ー類似度一覧表の例を示す図

【図８】正解分類情報格納例を示す図

【図９】登録文書データの削除手順を示す図

【図１０】正解／不正解影響度格納例を示す図

【図１１】正解／不正解影響度条件格納例を示す図

【図１２】削除文書候補格納例を示す図

【図１３】登録文書数分布範囲指定テーブル格納例を示す図

【図１４】登録文書格納例を示す図

【図１５】削除対象度合い算出及び格納例を示す図

【符号の説明】

１…制御装置

２…入力装置

３…表示装置

４…外部記憶装置

２００…プログラム部

２０１…初期化部

２０２…分類文書入力部

２０３…登録文書読み込み部

２０４…類似度算出部

２０５…分類特定部

２０６…正解分類情報入力部

２０７…正解／不正解影響度算出部

２０８…削除候補判断条件設定部

２０９…影響度判定・削除候補出力部

２１０…削除文書選択部

２１１…登録文書削除部

２１２…分類別登録文書数範囲設定部

２１３…分類別登録文書数読み込み部

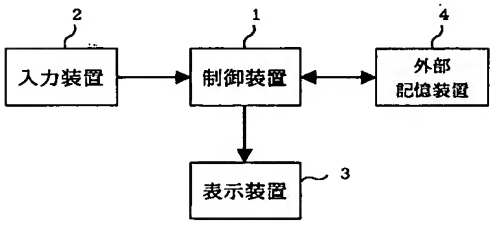
15

２５０…バッファ部
２５１…分類文書格納バッファ部
２５２…登録文書格納バッファ部
２５３…類似度算出結果格納バッファ部
２５４…正解分類情報格納バッファ部

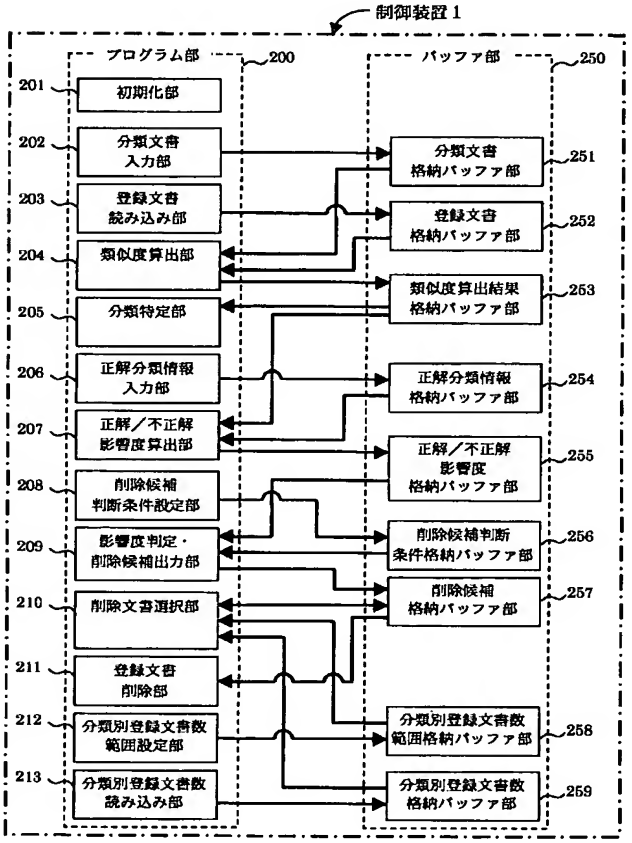
16

２５５…正解／不正解影響度格納バッファ部
２５６…削除候補判断条件格納バッファ部
２５７…削除候補格納バッファ部
２５８…分類別登録文書数範囲格納バッファ部
２５９…分類別登録文書数格納バッファ部

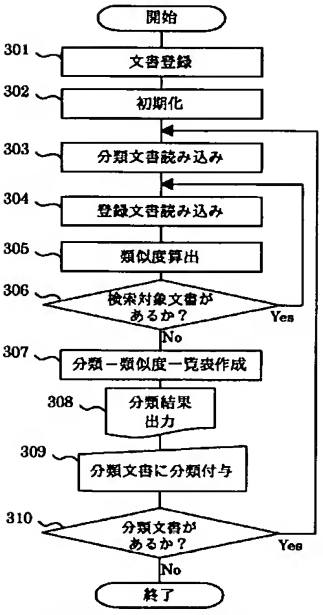
【図１】



【図２】



【図３】



（第１のステップ：分類処理）

【図４】

分類文書ID	1
本文	この文書は、パソコンについて記述したものです。

（分類文書格納例）

【図５】

登録文書ID	1
分類	印刷
本文	この文書は、印刷について記述したものです。
登録文書ID	2
分類	テレビ
本文	この文書は、テレビについて記述したものです。
	：

（登録文書格納例）

【図６】

分類文書ID	1	
登録文書ID	分類	類似度
1023	テレビ	0.378
9924	パソコン	0.226
5933	印刷	0.172
75268	テレビ	0.039
：	：	：

（類似度算出結果バッファ部格納例）

【図７】

分類	類似度の和
パソコン	1.782
ビデオ	1.023
テレビ	0.729
印刷	0.514
：	：

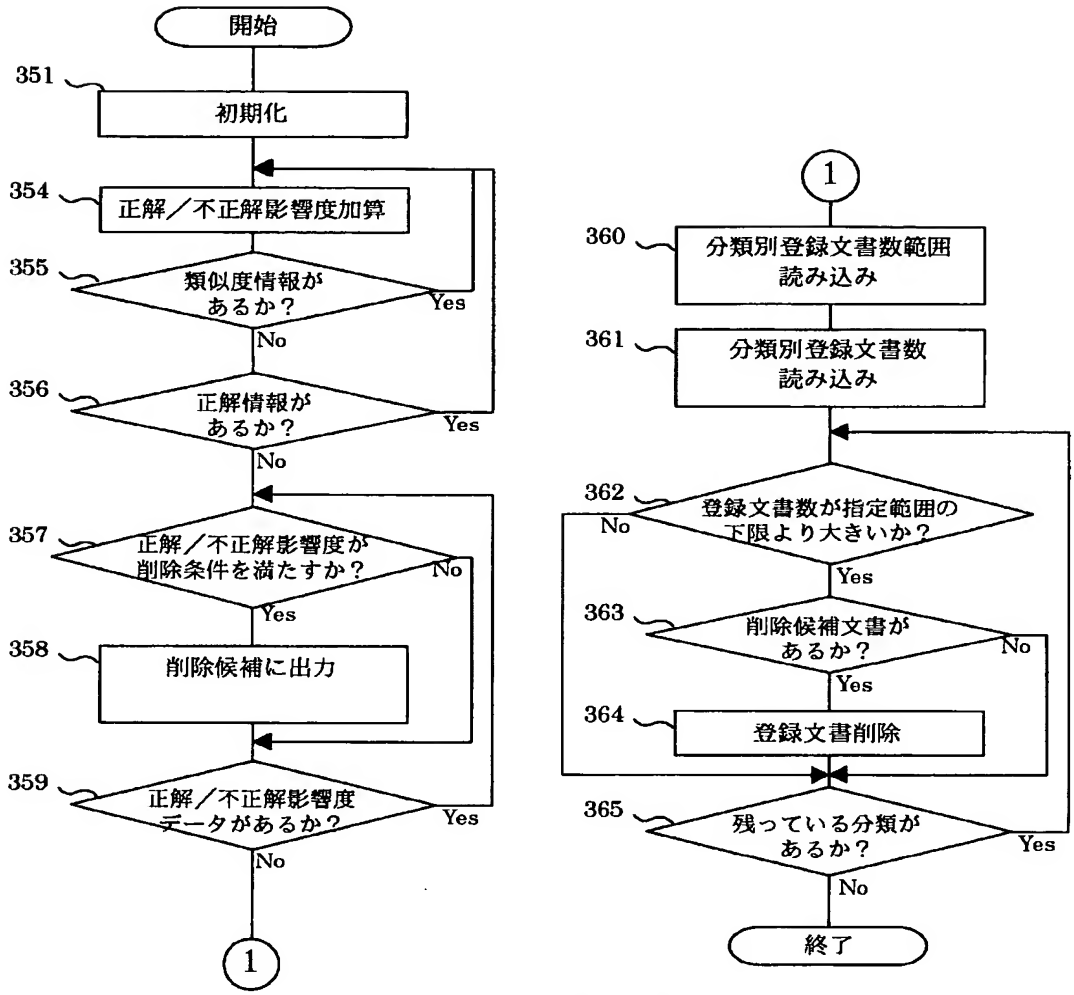
（分類－類似度一覧表例）

【図８】

分類文書ID	正解分類
1	パソコン

（正解分類情報格納例）

【図９】



（第２のステップ：登録文書削除処理）

【図１０】

登録文書 ID	分類	正解影響度	不正解影響度
1	家具	0.56285	0.00845
2	自動車	0.00000	0.23582
:	:	:	
1023	テレビ	0.08722	1.35160
:	:	:	
9924	パソコン	0.00021	0.35991
:	:	:	

（正解／不正解影響度格納例）

【図１１】

正解影響度条件	0.01 未満
不正解影響度条件	0.20 以上

（正解／不正解影響度条件格納例）

【図１３】

分類	分布範囲
パソコン	3±0.5%
インターネット	6.5±1.0%
:	:

（登録文書数分布範囲指定テーブル格納例）

【図１２】

登録文書 ID	分類	正解影響度	不正解影響度
2	自動車	0.00000	0.23582
9924	パソコン	0.00021	0.35991
:	:	:	

（削除文書候補格納例）

【図１４】

分類	文書数（比率）
全文書数	2806 件
パソコン	90 件（3.2%）
インターネット	119 件（4.2%）
:	:

（登録文書数格納例）

【図１５】

登録文書ID	分類	正解影響度	正解回数	不正解影響度	不正解回数	登録後月数	削除対象度合	順位
40	パソコン	0.53131	4	0.00353	1	15	0.150	4
321	パソコン	1.39884	7	0.00000	0	3	-1.927	6
2056	パソコン	0.01428	2	0.76452	3	10	0.952	1
6772	パソコン	1.14297	6	3.07939	8	2	0.714	2
7109	パソコン	0.00000	0	0.00000	0	8	0.080	5
9924	パソコン	0.00021	1	0.35991	4	9	0.275	3

（削除対象度合い算出及び格納例）

フロントページの続き

(72)発明者　齋藤　裕美
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内
(72)発明者　小林　勉
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内
(72)発明者　松隈　剛
東京都青梅市新町３丁目３番地の１　東芝コンピュータエンジニアリング株式会社内
(72)発明者　中本　幸夫
東京都青梅市新町３丁目３番地の１　東芝コンピュータエンジニアリング株式会社内
(72)発明者　仁科　卓哉
東京都青梅市新町３丁目３番地の１　東芝コンピュータエンジニアリング株式会社内
(72)発明者　山崎　弘
東京都青梅市新町３丁目３番地の１　東芝コンピュータエンジニアリング株式会社内

Ｆターム(参考)　5B075 ND03 NR03 NR12 PP02 PQ02 PR06 QM08